**Unsupervised Learning with Dimensionality Reduction and Clustering**

Samarjith D,

R V College Of Engineering, Section 1

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. **Abstract**

This project explores the application of **unsupervised learning techniques** to analyze **high-dimensional datasets**. The work was carried out in two phases using two different datasets.

In the **first phase**, the **MNIST handwritten digits dataset** was used. **K-Means clustering** was applied directly to the **64-dimensional digit data**, and the resulting cluster centers were found to resemble digit patterns, showing how unsupervised learning can capture meaningful structures even without labels. To improve interpretability and efficiency, **Principal Component Analysis (PCA)** was then applied to reduce the high-dimensional digit data to **two dimensions**, enabling clear visualization of clusters.

In the **second phase**, the focus shifted to the **Breast Cancer Wisconsin dataset** from scikit-learn, which contains **30 features** describing the cell nuclei of breast mass samples. The methodology involved **loading the dataset**, performing **feature scaling**, applying **PCA** to reduce the feature space to **two dimensions**, and then performing **K-Means clustering** to group the data into clusters. The quality of the resulting clusters was evaluated using the **silhouette score**, which indicated good separation and compactness of the clusters.

The findings from both datasets demonstrate that the **combination of PCA and K-Means** is highly effective in uncovering the underlying structure of data, successfully grouping samples into **distinct and interpretable clusters**. Visualizations using **Matplotlib** and **OpenCV** further reinforced the clarity of these results, highlighting how unsupervised learning can be leveraged for dimensionality reduction and pattern discovery in complex, high-dimensional datasets.

1. **Introduction**

**Unsupervised learning** is a key branch of machine learning where algorithms identify **patterns and structures** in **unlabeled data**. In many real-world scenarios, datasets are **high-dimensional**, making them difficult to analyze and visualize.

This project combines **dimensionality reduction** and **clustering** to address this challenge. **Principal Component Analysis (PCA)** is first applied to reduce complex datasets into **two dimensions**, enabling clear visualization. Then, **K-Means clustering** is used to group similar data points, revealing hidden structures without prior labels.

The work was carried out in two phases. In the first phase, the **MNIST handwritten digits dataset** was used. K-Means was applied to the 64-dimensional digit data, and PCA was used to reduce it to 2D for efficient visualization.

In the second phase, the **Breast Cancer Wisconsin dataset** (30 features) was analyzed. After scaling and applying PCA, K-Means clustering was performed. The quality of clusters was evaluated using the **Silhouette Score**, indicating strong separation and compactness.

The project was implemented in **Python** using libraries like **scikit-learn**, **NumPy**, **pandas**, and **Matplotlib**. During the initial two weeks of the internship, training covered:

* Python for Data Science (NumPy, pandas)
* Data Visualization (Matplotlib, Seaborn)
* Machine Learning Fundamentals: Supervised vs. Unsupervised Learning
* Clustering Algorithms (K-Means, Hierarchical)
* PCA and Dimensionality Reduction
* Evaluation Metrics like Silhouette Score

This project demonstrates how combining PCA with K-Means can **simplify high-dimensional data** and uncover **meaningful, interpretable patterns**, with applications in **bioinformatics**, **healthcare**, and market segmentation.

1. **Project Objective**

### Project Objectives

The primary objective of this project was to **implement and evaluate a complete unsupervised learning pipeline**. The specific goals are:

1. **Dataset Selection and EDA**: Select a suitable high-dimensional dataset (MNIST and Breast Cancer) and perform **exploratory data analysis** to understand its characteristics and distributions.
2. **Dimensionality Reduction with PCA**: Apply **Principal Component Analysis (PCA)** to reduce the dataset from its original feature space to **two principal components**, enabling easier visualization.
3. **Clustering with K-Means**: Implement the **K-Means algorithm** on the dimensionally-reduced data to partition it into a predefined number of clusters (**k=3** for the Breast Cancer dataset).
4. **Cluster Evaluation**: Quantitatively evaluate the quality and separation of clusters using the **Silhouette Score**, ensuring meaningful groupings.
5. **Cluster Visualization**: Visualize the resulting clusters and centroids on a **2D scatter plot** to qualitatively assess the clustering performance and interpret the underlying structure of the data.
6. **Methodology**

The methodology followed in this project involved a systematic application of **unsupervised learning techniques** on two high-dimensional datasets: **MNIST handwritten digits** and **Breast Cancer Wisconsin dataset**. The project workflow can be divided into the following steps:

### 1. Data Collection

* **MNIST Dataset**: This is a publicly available dataset of **70,000 handwritten digit images** (28x28 pixels), where each image is represented as a **64-dimensional feature vector**. It was directly loaded using scikit-learn’s load\_digits() function.
* **Breast Cancer Dataset**: This dataset contains **569 samples** with **30 numeric features** describing cell nuclei characteristics. It was loaded using scikit-learn’s load\_breast\_cancer() function.

### 2. Exploratory Data Analysis (EDA)

* Basic EDA was performed to **understand the dataset characteristics**, check for missing values, and examine distributions of features.
* Tools used: **pandas**, **NumPy**, **Matplotlib**, **Seaborn**.
* Steps included:  
  + Viewing **data shape** and **feature names**.
  + Checking for **missing or null values**.
  + Calculating **summary statistics** (mean, standard deviation, min, max).
  + Visualizing **feature correlations** using heatmaps.
  + Plotting **target distribution** for the Breast Cancer dataset.

### 3. Data Preprocessing

* **Standardization**: Features were scaled using **StandardScaler** to ensure that all features contribute equally to distance calculations in clustering.
* **Dimensionality Reduction (PCA)**:

PCA was applied to both datasets to reduce dimensionality:

**MNIST**: 64 → 2 components

**Breast Cancer**: 30 → 2 components

This step facilitated **visualization** and **computational efficiency**.

### 4. Clustering

* **Algorithm**: K-Means clustering was selected due to its simplicity, efficiency, and effectiveness for numeric high-dimensional data.
* **Model Implementation**:  
  + Number of clusters (**k**) was chosen based on dataset characteristics:  
    - MNIST: k=10 (digits 0–9)
    - Breast Cancer: k=3 (arbitrary for illustration; silhouette score used to validate)
  + The model was trained on the **PCA-reduced 2D data**.
* **Validation**:  
  + **Silhouette Score** was calculated to assess cluster quality (compactness and separation).

### 5. Visualization

* **Matplotlib**: Used to plot **2D scatter plots** showing clusters and centroids.
* **OpenCV**: Used to generate **graphical visualizations** of cluster points on a canvas, especially for a visual and interactive representation.
* PCA ensured that the **high-dimensional data** could be visualized clearly in 2D while preserving variance.

### 6. Tools and Technologies

* **Programming Language**: Python 3.x
* **Libraries Used**:  
  + scikit-learn (datasets, preprocessing, PCA, K-Means, silhouette\_score)
  + pandas and NumPy (data handling, preprocessing)
  + Matplotlib and Seaborn (visualizations)
  + OpenCV (cluster visualization)

### 7. Model Selection and Validation

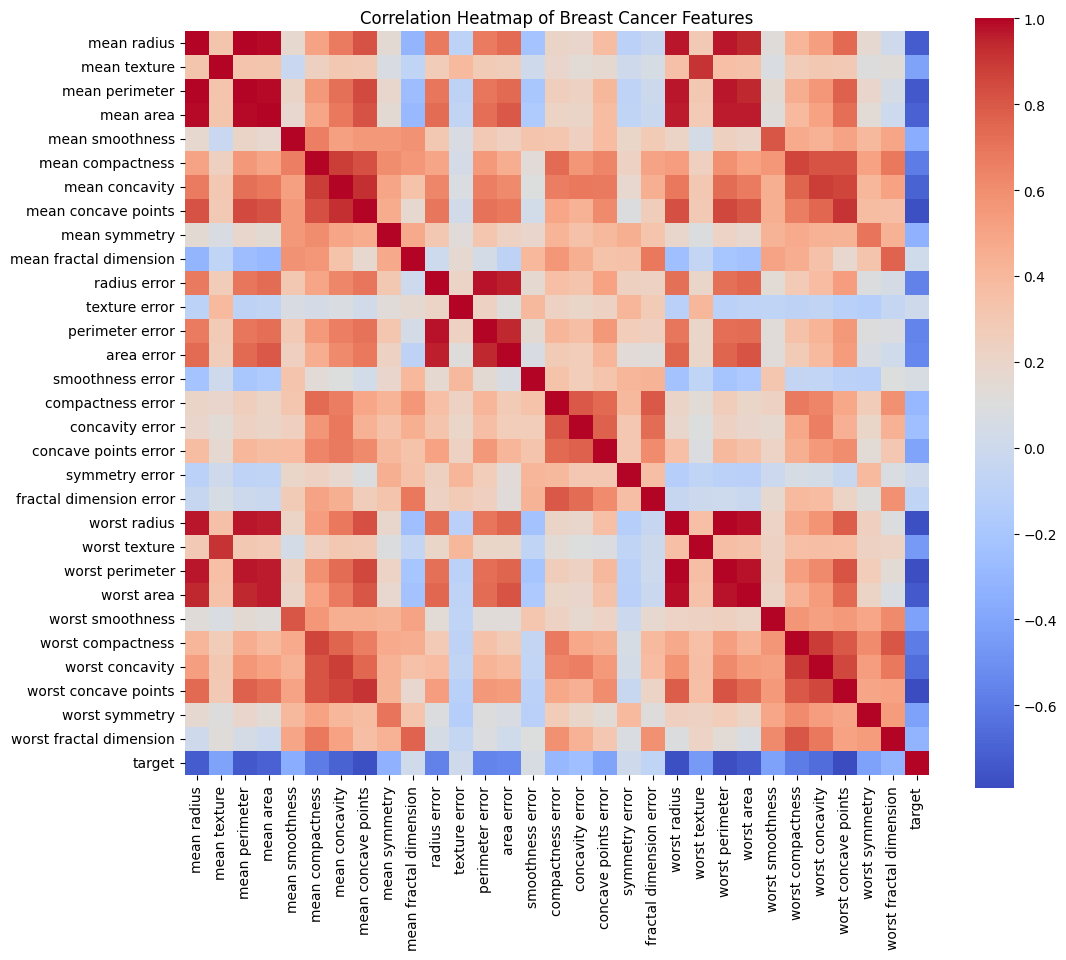
* **Model Choice**: K-Means was chosen for its efficiency with numeric datasets and ease of interpretation.
* **Validation Metric**: **Silhouette Score** was used to validate cluster quality. A higher score indicates better separation and compactness of clusters.
* **Training and Testing**: Since K-Means is **unsupervised**, the model was applied to the **entire dataset**. PCA was used to reduce dimensionality for both training and visualization purposes.

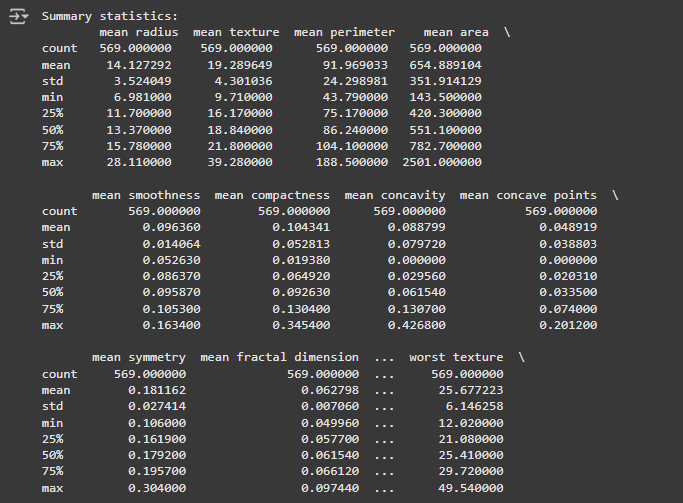
### 8. Code Repository

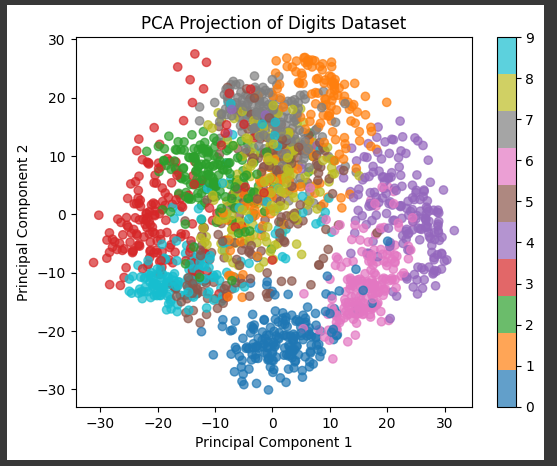
All the Python code for data loading, preprocessing, PCA, clustering, evaluation, and visualization is hosted on **GitHub**:

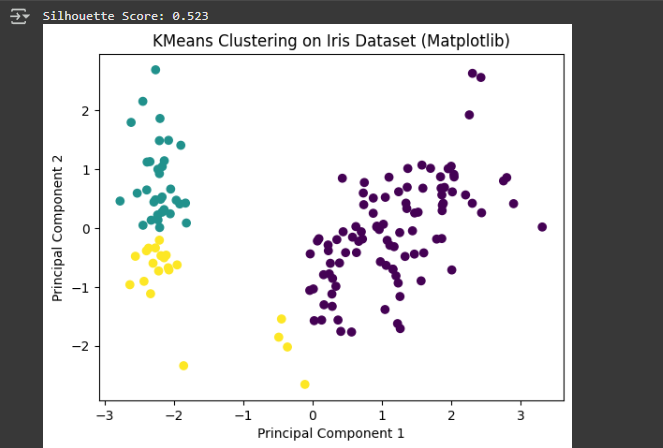
**GitHub Link:**https://github.com/samarjithd/IDEAS-TIH-Unsupervised-Learning

1. **Data Analysis and Results**









1. **Conclusion**

This project demonstrated the effective use of **unsupervised learning** to analyze high-dimensional datasets. By combining **PCA for dimensionality reduction** with **K-Means clustering**, we were able to uncover meaningful patterns in both the **MNIST handwritten digits** and **Breast Cancer Wisconsin** datasets. The **Silhouette Score** confirmed that the clusters were well-separated and compact, especially for the Breast Cancer dataset. Visualizations using **Matplotlib** and **OpenCV** provided clear, interpretable representations of the clusters. Overall, the project highlights how unsupervised learning can simplify complex datasets, reveal hidden structures, and support exploratory data analysis in real-world applications such as **healthcare and pattern recognition**.

1. **APPENDICES**

## References

1. **Scikit-learn Documentation** – https://scikit-learn.org/stable/
2. LeCun, Y., Cortes, C., & Burges, C. **MNIST Handwritten Digit Database** – http://yann.lecun.com/exdb/mnist/
3. Dua, D. & Graff, C. **UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set** – https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
4. Jolliffe, I. **Principal Component Analysis**, Springer Series in Statistics, 2002.
5. Xu, R., & Wunsch, D. **Clustering Algorithms in Data Mining: An Overview**, IEEE Transactions on Neural Networks, 2005.

## GitHub Link for Codes

* All Python codes for **data loading, preprocessing, PCA, K-Means clustering, evaluation, and visualization** are hosted on GitHub:  
  <https://github.com/samarjithd/IDEAS-TIH-Unsupervised-Learning>

## Other Document Links

* **Project Video Demo**:<https://drive.google.com/file/d/1bVpFkBqZybBFAt1PPgEtnihsQUg56xa4/view?usp=sharing>
* **Project Report, Data Sheet, and Presentation**: Available in the same GitHub repository listed above.